University of Vermont
ScholarWorks @ UVM

UVM Honors College Senior Theses

Undergraduate Theses

2015

Exploring the Evolution of Modularity in Gene Regulatory Networks

Mariko L. Totten
University of Vermont, mtotten@uvm.edu

Exploring the Evolution of Modularity

in Gene Regulatory Networks

An Undergraduate Honors College Thesis

Presented By:
Mariko Totten

Advised By:
Joshua Bongard, Ph. D.

Department of Computer Science
University of Vermont, Burlington, 05405

April 23, 2015

Contents

Abstract

Modularity has been observed in biological systems of all shapes and sizes. Intuitively, we recognize that this component-based structure can increase efficiency and robustness in a system. Gene regulatory networks are a perfect example of a biological system that exhibits modularity, however, just how this modularity evolved is still unresolved. The research of Espinosa-Soto and Wagner supports the hypothesis the modularity observed in biological gene regulatory networks evolved as a result of specialization in gene activity (Espinosa-Soto & Wagner, 2010). By manipulating the implementation of their research we may further explore the conditions that drive the evolution of modularity. Specifically, this study explores how the robustness of a network and use of biased mutation may influence how modularity is evolved in gene regulatory networks.

1. Introduction

This study investigates the evolution of modularity in gene regulatory networks (GRNs). GRNs dictate cell behavior and organism development and can be modeled using Boolean networks. GRNs have a modular structuring, but how this modularity came to evolve in the first place is still a topic of debate. Computer modeling can provide critical insight into the behavior of GRNs and could formulate hypotheses to probe how modularity evolves. Research performed by Carlos Espinosa-Soto and Andreas Wagner in their 2010 study "Specialization Can Drive the Evolution of Modularity" (Espinosa-Soto & Wagner, 2010) concluded that, as the title suggests, specialization in gene activity may cause modularity to evolve in GRNs. This research examined how certain aspects of the model used by Espinosa-Soto and Wagner (2010) influenced their findings on the evolution of modularity.

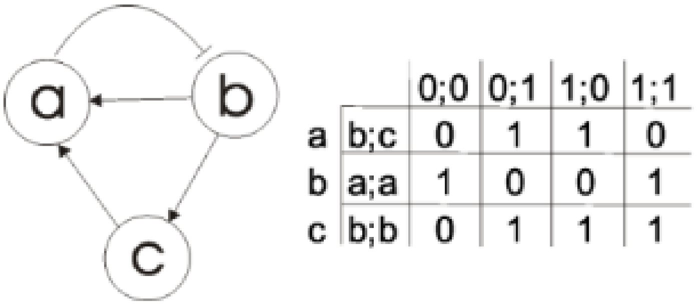
1.1 Modeling the GRN

While every cell in an organism contains the same set of genes the way individual cells behave varies extraordinarily. This is because a cell's behavior is defined by how the cell's genes are expressed. The process of changing a cell's gene expression through the activation and repression of genes is called gene regulation. Subsequently, the GRN, which is comprised generally of genes and the interactions between them, is the control system for organism development.

Gene activity pattern (GAP) refers to the collective activity states of all of the genes in a cell at a given time. It is this pattern that defines a cell's behavior and subsequently an organism's phenotype. In this way, the GAP can be viewed as a cell's identifying signature. For example, the gene regulation that occurs during development may cause stem cells to fall into two unique GAPs that distinguish which cells are to become liver cells and which will become heart cells. Gene regulation also occurs over the course of an organism's lifetime in response to environmental changes. In this case, external inputs may pressure a set of cells to adjust their current GAP, for example causing fur to grow thicker in colder weather.

As one can imagine, when this network is composed of thousands of inter-regulating genes the complexity of interactions and resulting behavior becomes difficult to follow. Biologists can research the actions of specific genes but difficulty studying the network as a whole. In order to predict the behavior of a network it becomes critical to look at the entire set of interactions rather than focusing on individual genes (Fernández & Solé, 2003). Mathematical models can be used to simulate complex interactions in order to emulate a GRN. Mathematically, a GRN can be understood as a mixed graph, G(N,U,D) in which the nodes, N, represent the network's genes and their associated activity. The set of directed edges, D, represent the direct causal interactions between genes, and the undirected edges, U, indicate associations between genes as a result of hidden confounding variables (Das, Caragea, Welch, & Hsu, 2010).

Specifically, researchers have used Boolean networks as a simplified model for observing and predicting the behavior of GRNs. "The biological basis for the development of Boolean networks as models of genetic regulatory networks lies in the fact that during regulation of functional states, the cell exhibits switch-like behavior," (Shmulevich, Dougherty, Kim, & Zhang, 2002). A Boolean network is comprised of a set of nodes with associated Boolean states. Nodes can be either on or off, which emulate genes in either the activated or repressed state. The state of each node is determined at every time step through a Boolean function, which uses logical functions to specify output for a given set of inputs (Fernández & Solé, 2003). In other words, the expression of each gene is determined at these time steps as a Boolean function of the current GAP (see Figure 1). Despite its obvious simplifications, Boolean networks have been



		0;0	0;1	1;0	1;1
a	b;c	0	1	1	0
b	a;a	1	0	0	1
c	b;b	0	1	1	1

Figure 1. The left image shows a graphical representation of a Boolean network with nodes a, b, and c, and their interactions indicated by edges. The truth table on the right describes the logical functions used to determine the output of each node (Speith, 2004).

shown to accurately predict the segmentation of Drosophila melanogaster in the work of Reka Albert and Hans Othmer (Albert & Othmer, 2003).

By applying computational evolution to these Boolean networks we can produce a model of how GRNs may evolve. Computational evolution refers to the concept of using computational algorithms to simulate the evolution of a population in order to produce a “more fit” population. In particular, the evolution of GRNs can be modeled using canonical genetic algorithms. These algorithms are characterized by two main components – selection and variation. First, the population must be assigned a goal by which its members will be evaluated. How well a member of the population is able to achieve this goal is called the member’s fitness. The level of fitness a member attains in relation to others in the population determines whether that member will be *selected* to seed the next generation. The best performing members at the end of the fitness evaluation then undergo *variation* through random mutation and/or sexual recombination in order to produce a new generation of population members. This process of selection and variation over many generations creates pressure to continually improve the average fitness of the population (Eiben, 2003). Evolutionary algorithms can be used to find non-intuitive solutions to a problem or, in the case of this research, examine topics related to the evolutionary process. Specifically this research investigates how modularity may be produced through evolution.

1.2 Modularity in the GRN

Modularity is the partitioning of a structure into semi-autonomous sub-units, and is ubiquitous in biology (Callebaut & Rasskin-Gutman, 2005; Schlosser & Wagner, 2004)) . Inherently we can understand the benefits of such a structuring, and modularity has been identified as one of the three critical components in the study of complex biological systems, along with robustness and emergence (Aderem, 2005). In order to understand a biological system it is critical to study not only how a system may be partitioned into modules, but what effect this has on the system and how this modular structuring came to evolve in the first place.

There are two outstanding characteristics of modules: "their *integration* concerning their internal relations (between their components) and their *autonomy* concerning their external relations (to elements of the context)" (Schlosser & Wagner, 2004, p. 4-5). Thus the behavior of a module relies strongly on the functioning of other elements within that module, but little on the elements outside of it. As a result of the simple characteristic of modularity, there are two critical benefits of a modularly structured system. First, modularity improves the robustness of a system, or the ability to maintain behavior while experiencing external perturbations. When perturbations cause changes to a single module, any problems that arise are restricted to that module with little effect on the entire system (Aderem, 2005). Additionally, connectivity within a module may help to rapidly and effectively correct for the perturbation.

Furthermore, modularity sets up that system for more efficient and effective evolution. Modularity improves evolvability through "the dissociation of developmental processes (e.g. heterochrony), the duplication of subsequent divergence of developmental modules, and the co-option of features into new functions" (Lorenz, Jeng, & Deem, 2011, p. 33). In simpler terms, modularity provides the ability for change to occur within one module without affecting the greater system, such that each module develops to perform a specific independent function. These functions can then change over time to become more effective or be slightly adjusted to be utilized in a different circumstance.

Modularity in a biological network refers to the idea that the network can be partitioned into densely interconnected groups of nodes with little intragroup connectivity (see Figure 2). Specifically in a GRN, this means that within a module the genes greatly influence the activity state of one another, but do not greatly affect the state of any genes outside of the module. Modularity in the regulatory network is especially noteworthy because “modularity in the phenotype is an immediate consequence of a developmental modularity,” (Lorenz et al., 2011, p. 33). Besides imposing modularity on the phenotype, modularity in the GRN also has been observed to foster robustness and evolvability. GRNs “maintain their intrinsic behavior even when they are perturbed externally,” (Espinosa-Soto & Wagner, 2010, 2) and make the “evolution of new complex circuits and resulting phenotypes easier,” (Lorenz et al., 2011, p.23).

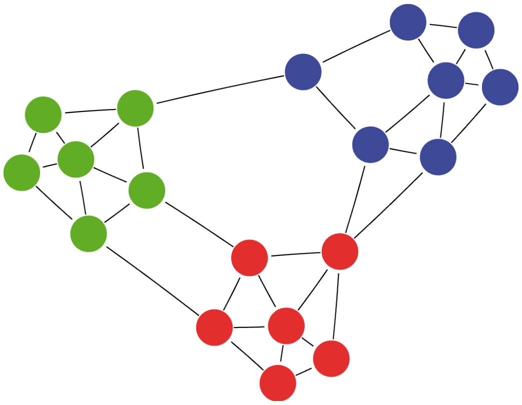


Figure 2. This graphical representation of a network demonstrates the partitioning of nodes into modules, indicated by the red, green and blue colored groups. Observing the edges between nodes shows dense connectivity within each module and sparse connectivity between modules (Amsen, 2013).

While modularity and its effects have been observed in GRNs, there is still no definitive understanding of how and to what strength this modularity evolves. Previous research suggested three prominent theories regarding why modularity evolves: 1) Modularity arises from a pressure to reduce the connection costs of interactions between nodes (Clune, Mouret, & Lipson, 2013); 2) Modularity occurs as a result of simultaneous directional and stabilizing selection (Wagner, 1996); 3) Modularly-varying evolutionary goals due to changes in organisms' environment drives the evolution of modularity (Kashtan & Alon, 2005).

1.3 Espinosa-Soto and the Evolution of Modularity in GRNs

A study by Espinosa-Soto and Wagner provided an alternative hypothesis for the evolution of modularity in GRNs (Espinosa-Soto & Wagner, 2010). Their work used evolutionary computation on a Boolean model of GRNs to support their hypothesis that specialization in gene activity drives the regulatory network toward a modular organization.

Specialization is the creation of a new GAP that may occur during a lifetime, such as alteration of a specific body function due to certain environmental conditions. For example, consider the gene regulation that occurs during organism development. The environment surrounding a stem cell signals whether that cell should become a liver cell rather than a heart cell. Subsequently, some portion of the GAP is defined to exhibit the unique characteristics of a liver cell. Alternatively, other traits are consistent throughout many different types of cells in the organism despite different environmental conditions. For example, in all cells the regulatory network should produce the portion of the GAP that builds cell walls. The favoring of a ubiquitous trait such as this is referred to as generalization. Consequentially, selection must simultaneously favor maintaining the activity states of some set of genes while changing the desired activity states of others. Espinosa-Soto and Wagner (2010) suggest that modularity arises within the network so that both of these patterns can be accommodated simultaneously. Espinosa-Soto and Wagner's evolutionary model (2010) supported their hypothesis that modularity develops in GRN's as a result of specialization (see Figure 3.)

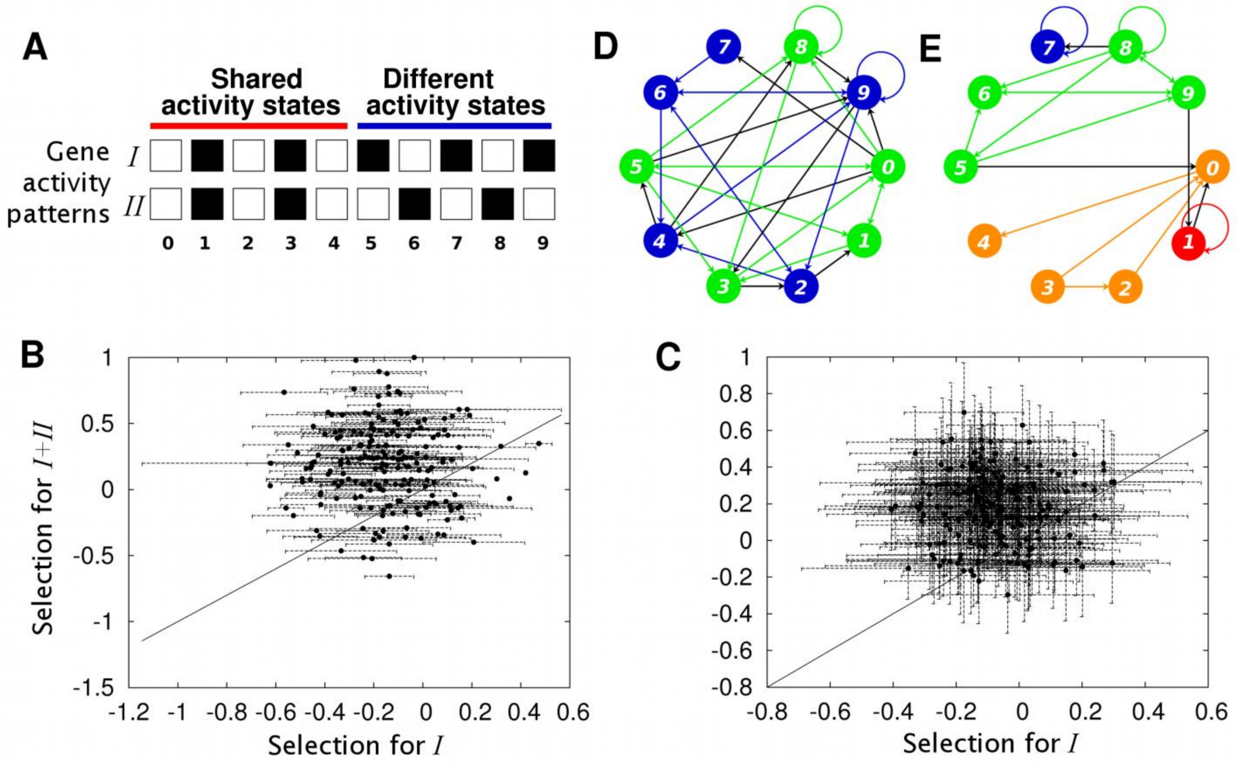


Figure 3. These figures show the results of the research performed by Espinosa-Soto and Wagner (2010). (A) shows the GAPs used their evolutionary simulations. Genes 0-4 in either GAP share the same activation states, to represent the ubiquitous expression of a trait. The states of genes 5-9 differ between GAPs to indicate specialization in gene activity unique to certain regions of the body or environmental conditions. (B) and (C) show the level of modularity, as measured by a normalized Q value (see equation 3), after selection for GAP I versus GAP I and II. Specifically (B) shows the modularity of the highest performing network in each population and (C) shows the average modularity of a population. In both graphs we see that modularity tends to be higher after selection for both GAPs. (D) and (E) are examples of high fitness networks, with modules indicated by different colors. (D) is a network after selecting for GAP I and (E) is after selection for I and II. Clearly, (E) is far more modular than (D).

1.4 Further Exploration of the Evolution of Modularity in GRNs

The research performed by Espinosa-Soto and Wagner (2010) provides a framework to continue exploring the evolution of modularity in GRNs. By recreating the basis of their research we can then manipulate aspects of their implementation to examine how these conditions may affect how strongly modularity evolves. This research seeks to explore two critical aspects of their implementation.

First, the equation for random mutation used by Espinosa-Soto and Wagner (2010) pressured networks to be sparsely connected, averaging two to three connections per gene. The researchers explained that this bias was used because it is an accurate reflection of the low connectivity observed in natural GRNs, among other biological networks. While this is true, the logic behind these specific numbers lies in the property of criticality in GRNs (Torres-Sosa, Huang, & Aldana, 2012). This property is a result of how the average number of incoming connections into each node of a network influences the average cycle length of the attractor a network falls into. Criticality refers to the distinct number of regulators that keeps this cycle length relatively small, around 1, but still allows the network to experience variation. Research has shown that in the right circumstances this level of criticality tends to be reached by two to three regulators per gene (Espinosa-Soto & Wagner, 2010). Additionally, by pressuring networks to remain sparse, this model essentially applies connection costs to the GRN. Recall, one theory for the evolution of modularity suggested that connection costs of interactions pressures networks to become modular (Clune et al., 2013). It is expected that removing this biased mutation will remove 'connection costs' and allow more regulators per gene, thus pushing the network out of criticality, and subsequently will decrease the level of modularity in a specialized GRN.

Additionally, Espinosa-Soto and Wagner's evolutionary model (2010) favored very robust networks. Networks were evaluated 500 times in the face of different perturbations, so that only the most robust networks received high fitness. However, as noted before, there is a significant relationship between modularity and robustness (Aderem, 2005). Studies suggest that modularity creates robustness in a network by isolating problems within a module and strengthening the ability of a module to correct itself. In turn it is conceivable that robustness may reinforce modularity, because a network that does not face many perturbations does not need to count on modules to confine problems. Therefore, it is informative to examine different levels of perturbations, which will in turn create selection pressure for varyingly robust networks, to observe how this may influence the evolution of modularity in specialized GRNs.

2. Methods

The research was conducted so as to replicate the 2010 research of Espinosa-Soto and Wagner's investigation of the evolution of modularity through specialization. Their methods were re-implemented as closely as possible given computational power and time restraints. Several specific variables were then modified to examine their role in the evolution of modularity.

2.1 Network Model

For this research, the GRN was modeled using a Boolean network. Though this model is a very simplified representation of a complex system, it can be argued that a Boolean network effectively captures the computational nature of gene regulation (Fernández & Solé, 2003). The network is comprised of ten genes each of which can be in either in an active (on) or repressed (off) state. The composite state of the genes at a given time defines a GAP, which is represented by a vector $\mathbf{s_t} = [s_t^0 , s_t^1 , \dots , s_t^9]$. An element s_t^i of this vector corresponds to the expression of gene i at time t, and can hold a value of either -1 (repressed) or 1 (active) (see Figure 4).

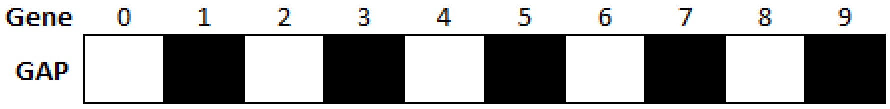
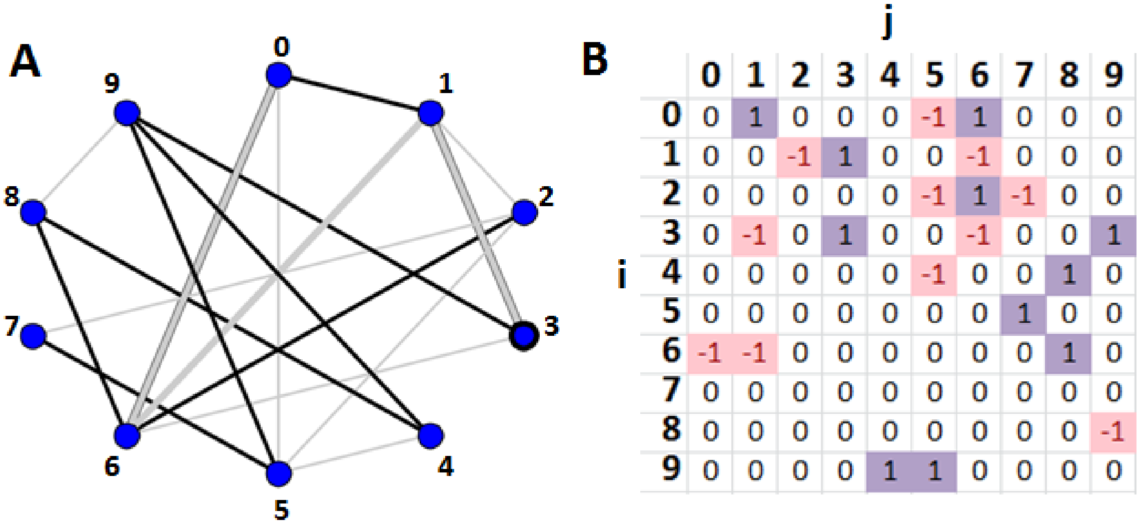


Figure 4. In this example of a GAP in the 10 gene network each block represents the activation state of its associated gene. White blocks indicate an activated gene and black blocks indicate a repressed gene.

The GAP at time t is determined by the interactions between the network's genes (see Figure 5A). These interactions are represented by a 10x10 adjacency matrix, $\mathbf{A} = [a_{ij}]$ (See Figure 5B). An element a_{ij} of this matrix signifies the influence gene j is exerting on gene i in the form of the activation ($a_{ij} = 1$), repression ($a_{ij} = -1$), or no interaction ($a_{ij} = 0$). Using these values, the state of a gene at the next time step can be determined by multiplying the current state of the gene by its regulators, as described in the formula:

$$s^{j}_{t+\tau} = \sigma\, [\, \Sigma^{10}_{j=1}\, a_{ij} s^{j}_{t}\,] \quad (1)$$

where $\sigma(x)$ equals 1 if $x > 0$ or -1 otherwise.



i \ j	0	1	2	3	4	5	6	7	8	9
0	0	1	0	0	0	-1	1	0	0	0
1	0	0	-1	1	0	0	-1	0	0	0
2	0	0	0	0	0	-1	1	-1	0	0
3	0	-1	0	1	0	0	-1	0	0	1
4	0	0	0	0	0	-1	0	0	1	0
5	0	0	0	0	0	0	0	1	0	0
6	-1	-1	0	0	0	0	0	0	1	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	-1
9	0	0	0	0	1	1	0	0	0	0

Figure 5. (A) shows a network of genes and their associated interactions. Thicker edged indicate that an interaction exists both from gene j to gene i and from gene i to gene j. Grey edges indicate a repressive influence and black edges indicate an activating influence. Bolded nodes indicate the existence of a self-loop, that is that gene i influences itself. (B) is the associated A_{ij} adjacency matrix for this network, where a_{ij} indicates the influence of gene j on gene i.

The GAP of a network dictates the expression of phenotypic traits. Specifically, in this model we consider the resulting phenotypic trait to be defined by the point attractors that the GAP falls into after a number of time steps. This model updates a network using the equation defined above until the network falls into one of two forms of attractors: fixed point, in which the same GAP is sustained indefinitely after a certain time step, or a cyclic attractor which repeatedly produces the same sequence of GAPs.

2.2 Simulating Specialization

Recall, specialization in GAPs refers to the idea that over a lifetime new GAPs may evolve in a specific body part or under certain environmental conditions (Espinosa-Soto &

Wagner, 2010). In order to demonstrate how specialization increases modularity, networks were first evolved without specialization. These networks were evaluated on their ability to attain GAP I. After these networks were successfully evolved, specialization was simulated by introducing a second pattern, GAP II. Networks were then evaluated on their ability to attain both pattern I and II. These patterns were specifically chosen so that patterns I and II shared the activity state of the first 5 genes (genes 0-4), and differed in activity state of the last 5 genes (genes 5-9) (see Figure 6).

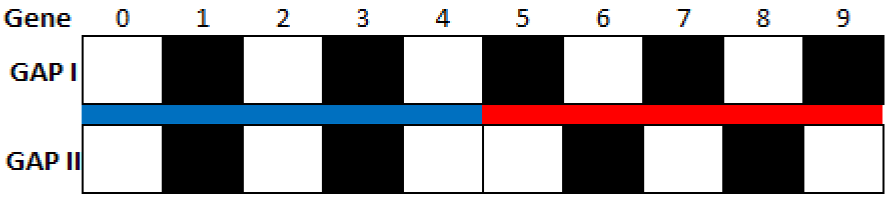
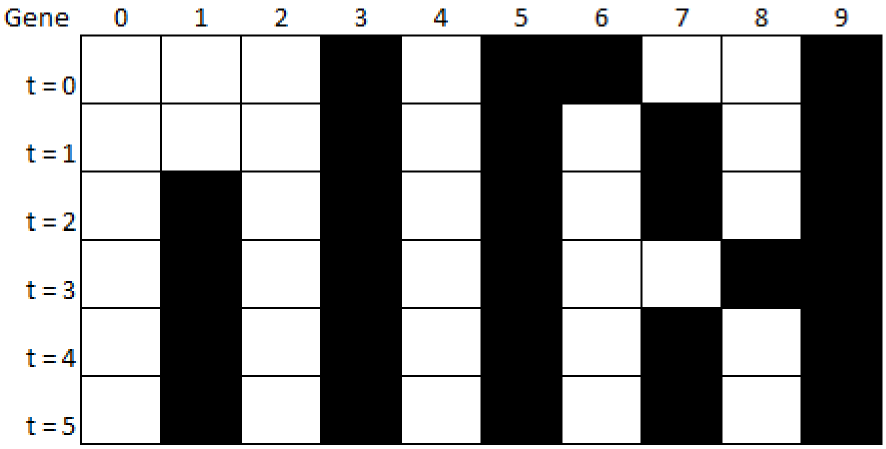


Figure 6. Networks are evolved to demonstrate either just GAP I or both GAP I and II. Black and white boxes represent inactive (s_i = -1) and active (s_i = 1) genes, respectively. The first five genes, indicated by the blue region, share the same activation in both GAPs. The latter 5 genes, indicated by the red region, differ in activation states

Espinosa-Soto and Wagner (2010) suggested that this specialization increases modularity because a network was most successful if it could maintain the state of the first five genes but change the states of the last five when facing different initial conditions. Thus a network would benefit by breaking interactions between these two sets of genes, creating two independently functioning modules.

2.3 Network Evaluations and Fitness

The fitness of each network is based on its ability to produce a fixed-point attractor for a specified GAP. Each network was evaluated using the following steps. First, the network's gene activity state is initialized to a perturbation of the desired GAP. These perturbations were created by altering the state of each gene in the desired GAP using a probability of p = 0.15. Then, beginning with this initial perturbation, the GAP was calculated at every time step using equation (1) until the network fell into either a point or cyclic attractor. Cyclic attractors were given a fitness of 0, because the goal of the network is to stably attain a GAP to simulate the stable expression of a phenotypic trait. Those networks that fell into point attractors were given a fitness value based on the Hamming distance, D, between the desired GAP and the final GAP of the system once it had fallen into a point attractor. Specifically, for each perturbation, a network was assigned a value of $\gamma = (1\text{-}D/D_{max})^5$, with accordance to Espinosa-Soto and Wagner (2010). Thus, best performing networks were those that were able to be initialized with a perturbation of a desired GAP, then remedy this perturbation so that the network fell into a point attractor that expressed the desired trait or GAP (see Figure 7).



Gene	0	1	2	3	4	5	6	7	8	9
t = 0	□	□	□	■	□	■	■	□	□	■
t = 1	□	□	□	■	□	■	□	■	□	■
t = 2	□	■	□	■	□	■	□	■	□	■
t = 3	□	■	□	■	□	■	□	□	■	■
t = 4	□	■	□	■	□	■	□	■	□	■
t = 5	□	■	□	■	□	■	□	■	□	■

Figure 7. This figure demonstrates the process of a successful network for GAP I. At t = 0 the network's genes are initialized with the perturbation of GAP I (genes 1, 6, and 7 have been switched). Equation (1) is then used to produce the new activation states at each time step. At t = 4 and t = 5 the exact same GAP is expressed, indicating that the network has fallen into a point attractor. The final activation state of each gene matches that of GAP I, so the Hamming distance is 0 and this network, at least with this perturbation, has performed well.

In order for a network to be robust it needed to be able to produce the desired GAPs in the face of many perturbations. The level of robustness in a network directly correlates to the number of perturbations the network experiences during evaluation. Espinosa-Soto and Wagner (2010) sought to examine only very robust networks, so all networks were evaluated against $\pi = 500$ perturbations. For this research, networks were evaluated against π = 20, 30, 40, 50, 75, or 100 perturbations in order to examine how the level of robustness influenced the strength of modularity that evolved.

Thus, this process of initializing a network with a perturbation and then updating until an attractor was reached was repeated for the assigned number of perturbations, π. Then the mean of the γ values for each of the π perturbations was calculated, $g = (\Sigma^{\pi}{}_{i=1} \gamma) / \pi$. Finally, the network fitness was found using the equation

$$\omega = f(g) = 1 - e^{-3g} \quad \textbf{(2)}$$

This fitness function was used in order to replicate the work of Espinosa-Soto and Wagner (2010).

2.4 Modularity

Modularity was measured using by calculating Q values to identify the strength of inter-module versus intra-module connectivity. For our model, we assumed that modularity would favor the partitioning of the network into two modules. These modules would be separated into genes that shared activity states in both GAPs (genes 0-4) and those that had opposing gene activity states between GAPs (genes 5-9). Specifically, Q is measured using the equation:

$$Q = \Sigma^{1}{}_{i=0} (l_i/L - (d_i/2L)^2) \quad \textbf{(3)}$$

In this equation, i represents the 2 modules that were identified above. L indicates the total number of edges in the network, l_i stands for the number of edges in module i, and d_i represents the total number of edges that leave, and enter, module i. This equation for modularity is a non-normalized substitute for the Q measurement used by Espinosa-Soto and Wagner (2010).

2.5 Evolutionary Simulations

Evolution was performed on populations of 100 networks. The initial population was created by randomly initializing a network with 20 interactions and then creating 99 mutations of this network. The mutation process was performed one of two ways.

First, to evaluate the varying levels of perturbations, mutation was performed as per the research of Espinosa-Soto and Wagner (2010). In this case, networks were mutated using a biased mutation operator that favored networks with low connectivity, about 2-3 regulators per gene. For each gene, the probability of a mutation occurring (μ) was 0.05. Given that a gene was to undergo mutation, the probability of losing an interaction was calculated

$$p(u) = (4r_u) / (4r_u + (N - r_u)) \quad \textbf{(4)}$$

In this equation, r_u represents the number of regulators on gene u and N is the number of genes in a network i.e. the maximum possible number of regulators on a gene. Conversely, the probability of acquiring an interaction was calculated using the equation $q(u) = 1 - p(u)$. A gene that was selected to acquire a new interaction had equal probability of gaining an activating or repressing regulator.

A second goal of this research was to investigate the influence of biased mutation on the evolution of modularity. In order to create control data to compare to that of biased mutation, evolution was performed using the simplest form of mutation. This completely unbiased mutation chose a single interaction at random from a network and then reassigned it a new random interaction. The new interaction had equal probability of flipping the state of the initial interaction value, removing the interaction, or remaining the same. This unbiased mutation was applied to populations using 75 perturbations, the minimal level of robustness found in this research to successfully recreate the results of Espinosa-Soto and Wagner (2010).

Once the initial population was created, networks were evaluated as described to attain fitness values, and the next generation was created using fitness proportional selection. That is, each network in the current generation was assigned a probability of selection based on the ratio of its fitness performance to the cumulative performance of the entire population, $P(i) = \omega_i / (\Sigma_j \omega_j)$. These probability ratios were then used to choose which networks would contribute to the next generation, such that those networks that performed better were more likely to be selected. Every network that was selected to contribute to the next generation then went through the appropriate mutation process as described above.

Populations were evolved for 500 generations, at which point networks could stably attain GAP I. Populations were then evolved for another 1500 generations and were evaluated for their ability to generate both GAP I and II. Q values were measured at the 500th and 2000th generation. A total of 20 evolutionary runs were performed for each level of perturbations. 20 additional runs were performed with the unbiased mutator for 75 perturbations.

2.6 Statistical Analysis

For each evolutionary run we examined the Q values for the networks with the highest fitness at 500 generations (after evolving for GAP I) and 2000 generations (after evolving for GAP I and II). The highest fitness in a population can often be achieved by multiple networks in that population. In some instances this is because the same network was selected from the previous generation and was not changed during mutation. For populations with multiple most-fit networks, duplicate networks were thrown out and then the mean Q value of the remaining best networks was calculated as a representative of the Q value of the most-fit network.

Thus for a given level of perturbation, each of the 20 evolutionary runs was assigned two Q values, one for the most-fit network after 500 generations and the other for the most fit network after 2000 generations. A one-tailed Student's T test was performed with an α of 0.05 to evaluate if the mean Q value after 2000 generations was significantly higher than the mean Q after 500 generations. The null hypothesis (h_0^x) was that $\mu_{II} <= \mu_I$ and the alternative hypothesis (h_1^x) was that $\mu_{II} > \mu_I$ where μ_I was the mean Q of the most fit network after evolving for only GAP I (500 generations), μ_{II} was the mean Q of the most fit network after evolving for both GAP I and II (2000 generations) and x was the number of perturbations used in the network evaluations.

3. Results

Examination of the mean modularity of high-performing networks after evolving populations for just GAP I and then both GAP I and II provides insight into the extent that modularity increases due to specialization.

3.1 Robustness and Modularity in Specialized GRNs

This research investigated how the level of robustness, as measured by the number of perturbations used to evaluate a network, influenced the evolution of modularity in a specialized GRN. For more robust networks evaluated using 75 and 100 perturbations we are 95% confident that we can reject the null hypothesis that the mean level of modularity, as measured by Q, of the highest performing networks after evaluating for GAP I and GAP II was equal to or less than the mean modularity after evaluation for GAP I alone. Alternatively, for an alpha of 0.05 the analysis did not suggest that we can reject the null hypothesis for less robust networks with any perturbation level lower than 75 (π = 20, 30, 40, 50) (see Table 1).

For populations that were evaluated with 75 perturbations the mean modularity increased by 17.2% between evolution for GAP I and evolution for GAP I and II, an increase of 0.057 from 0.328 to 0.385. This produced a t statistic of 1.829 and a corresponding p-value of 0.038. Populations evaluated with 100 perturbations increased 17.3% from 0.355 to 0.417, with a t-statistic of 1.731 and p-value of 0.046. Since in both these instances $p < 0.05$, this data suggests that the mean Q value of the fittest networks after introducing specialization in GAPs is significantly higher than without specialization.

The mean Q value also increased in population evaluated with 20, 30, 40, and 50 perturbations, but not enough to be considered statistically significant. The p-values for these means, in order of increasing number of perturbations, were 0.398, 0.061, 0.093, and 0.200.

Number of Perturbations (π)	20	30	40	50	75	100
Mean Q of Fittest Networks - GAP I	0.367	0.374	0.354	0.349	0.328	0.355
Mean Q of Fittest Networks - GAP I & II	0.374	0.423	0.390	0.374	0.385	0.417
T-Statistic	0.259	1.578	1.354	0.851	1.829	1.731
P-Value (One-Sided Student's T Test)	0.398	0.061	0.093	0.200	0.038	0.046

Table 1.Statistical analysis of mean modularity of the fittest networks in a population before and after the introduction of specialization in gene activity for varying levels of robustness.

3.2 Biased Mutation in the Evolution of Modularity in GRNs

The second part of this research investigated the effect of biased mutation on the evolution of GRNs. π = 75 perturbations were used since this level of robustness demonstrated a significant increase in modularity after specialization. When all bias was removed from mutation there was no statistical evidence that the level of modularity increased after evolving for GAP I & II. In fact, the mean Q value of the fittest networks decreased by 1%, producing a t statistic of -0.384 and a one-sided p-value of 0.648 (see Figure 9.)

Type of Mutation (π = 75)	Biased Mutation	Simple Mutation
Mean Q of Fittest Networks - GAP I	0.328	0.377
Mean Q of Fittest Networks - GAP I & II	0.385	0.373
T-Statistic (38 dof)	1.829	-0.384
P-Value (One-Sided Student's T Test)	0.038	0.648

Table 2. Statistical analysis of mean modularity of the fittest networks in a population before and after the introduction of specialization in gene activity with and without mutation bias.

4. Discussion

4.1 Implications

This research examined how the level of modularity evolved in GRNs, the dependent variable, was influenced by two different independent variables—network robustness and criticality.

4.1a Influence of Robustness on the Evolution of Modularity in Specialized GRNs. The examination of varying levels of perturbation on evolving modularity found statistical evidence that modularity was evolved in networks evaluated with at least 75 to 100 perturbations but not in networks with 20, 30, 40, or 50 perturbations. These results suggest that increasing the robustness of a network helps drive the evolution of modularity. However in order to examine the relationship between robustness and modularity it is critical to first examine the underlying relationship between robustness and fitness.

Recall that fitness is calculated by averaging performance over π perturbations of the desired GAP, where the probability that a gene state will be altered is 0.15. Consider first the scenario where a network is evaluated only once against an unperturbed copy of the desired GAP (we will call this $\pi = 0$). In this case it is not difficult for a network to find a solution since it does not need to be at all robust to change; a network that is able to obtain the desired GAP in one generation will be just as successful in the next generation (see Figure 8A). Now consider $\pi = 10$; some networks may get lucky and encounter no perturbations at one generation, receive a high fitness, and subsequently get copied to the next generation. However in the next generation it encounters many perturbations, a totally new experience for this network, so its fitness drops. In this case, it may be common for some networks in a population to perform very well after only a few generations, however, the spread of fitness across a population is very large because a network’s fitness may vary a lot between generations depending on the perturbations it encounters (see Figure 8B). Finally, consider the case where $\pi = 100$. Now, achieving a high fitness is not so trivial because a network has to be able to accommodate a variety of different conditions. However, this network, if not mutated, is less likely to experience a large change in fitness from one generation to the next because it has already seen a variety of perturbations. So, as we see in Figure 8C, it takes longer to increase fitness but the spread of population fitness decreases.

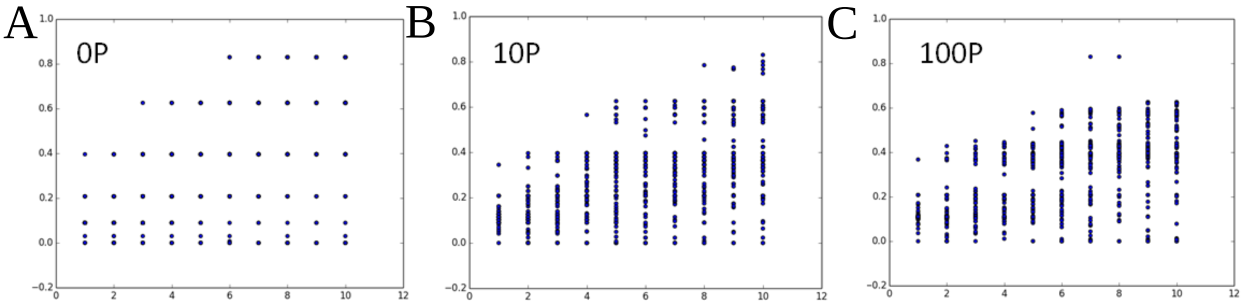


Figure 8. Each of these graphs demonstrate the fitness (y-axis) of each of 100 networks in a population over the first 10 generations of evolution (x-axis) using the number of perturbations indicated in the upper left corner.

This scenario displays the correlation between the robustness and fitness of our model GRNs. Increasing the number of perturbations that a network is evaluated against increases the evolutionary pressure for the population to produce robust networks. Subsequently, the more robust a population is, the more difficult it is to produce a network with very high fitness but also the more consistently networks behave from generation to generation.

It is important to note the differences in the fitness landscape as the number of perturbations increase because we need to ensure that the differences in statistical significance of modularity noted at different levels of perturbations are in fact a result of changes in robustness, not in the underlying effects on fitness. Specifically, in order to infer any significance it is necessary to evolve networks to a reasonably high level of fitness before analyzing Q values. Since it takes longer for more robust networks to achieve a high fitness it is critical that at all levels of perturbation evolution is carried out for enough generations for the population to produce networks of a reasonably high fitness, for evaluation of both GAP I and GAP I and II. Espinosa-Soto and Wagner (2010) designed their research to evolve highly robust networks, using π = 500 perturbations, and were able to obtain significant results by evaluating for GAP I for 500 generations and GAP I and II for an additional 1500 generations. Since this was a far greater level of robustness than performed in any of this research, it is assumed that using these same constraints provided enough time for the full evolution of all levels of perturbations. Additionally, Figure 9 demonstrates that there is no obvious difference in the fitness landscape of the fittest networks between varying levels of perturbations. While the effects of changing levels of robustness was obvious when examining populations for the first few generations, these differences were essentially smoothed out over many generations with the chosen numbers of perturbations.

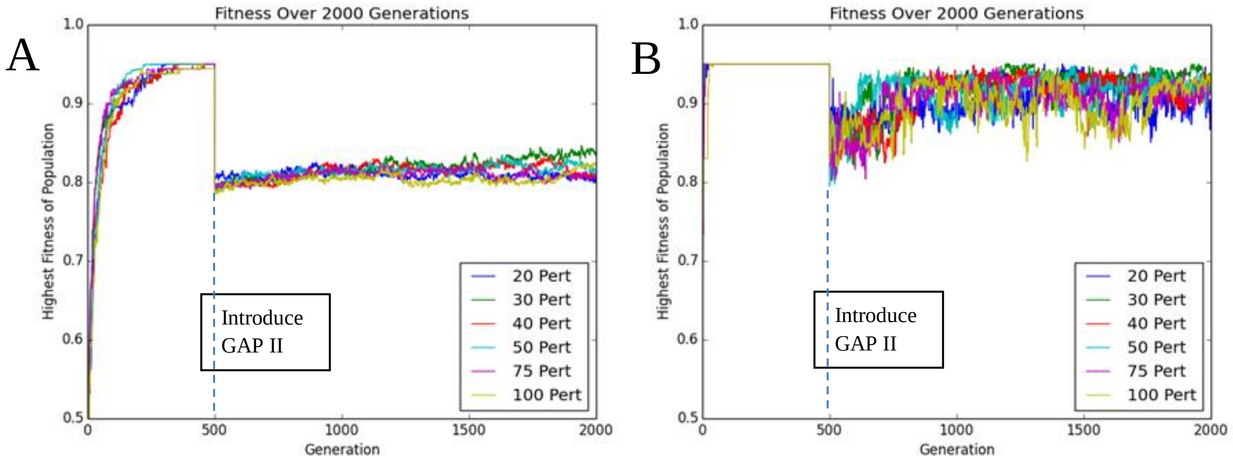


Figure 9. These graphs show the fitness of the best performing networks over 2000 generations. (A) gives the mean fitness of all 20 runs of the best performing network at each generation. (B) shows the fitness of the best performing network out of all 20 runs at each generation.

Given that there are no obvious differences in the fitness landscapes over 2000 generations between different levels of perturbations we can assume that it is in fact the variations in robustness that cause modularity to evolve differently. The fact that modularity only increased significantly when networks were evolved to be robust, π = 75 and π = 100, this suggests that robustness plays an important role in driving the evolution of modularity (see Figure 10). Recall that modularity is known to increase robustness in two ways. First, when a perturbation occurs its undesirable effects are isolated within a module due to sparse connectivity between modules. Secondly, within the module, dense connectivity is able to correct for the perturbation in the next time steps. By applying selection pressure for robustness in evolutionary simulations pressure is put on the evolving networks to discover modularity in order to solve problems created by perturbations.

Specifically, this research identified the threshold of robustness that is necessary in order for the expected evolution of modularity to occur. A statistically significant increase in modularity occurred only at 75 perturbations or above. Networks evolved with lower levels of perturbations found solutions that were far less modular (see Figure 11).

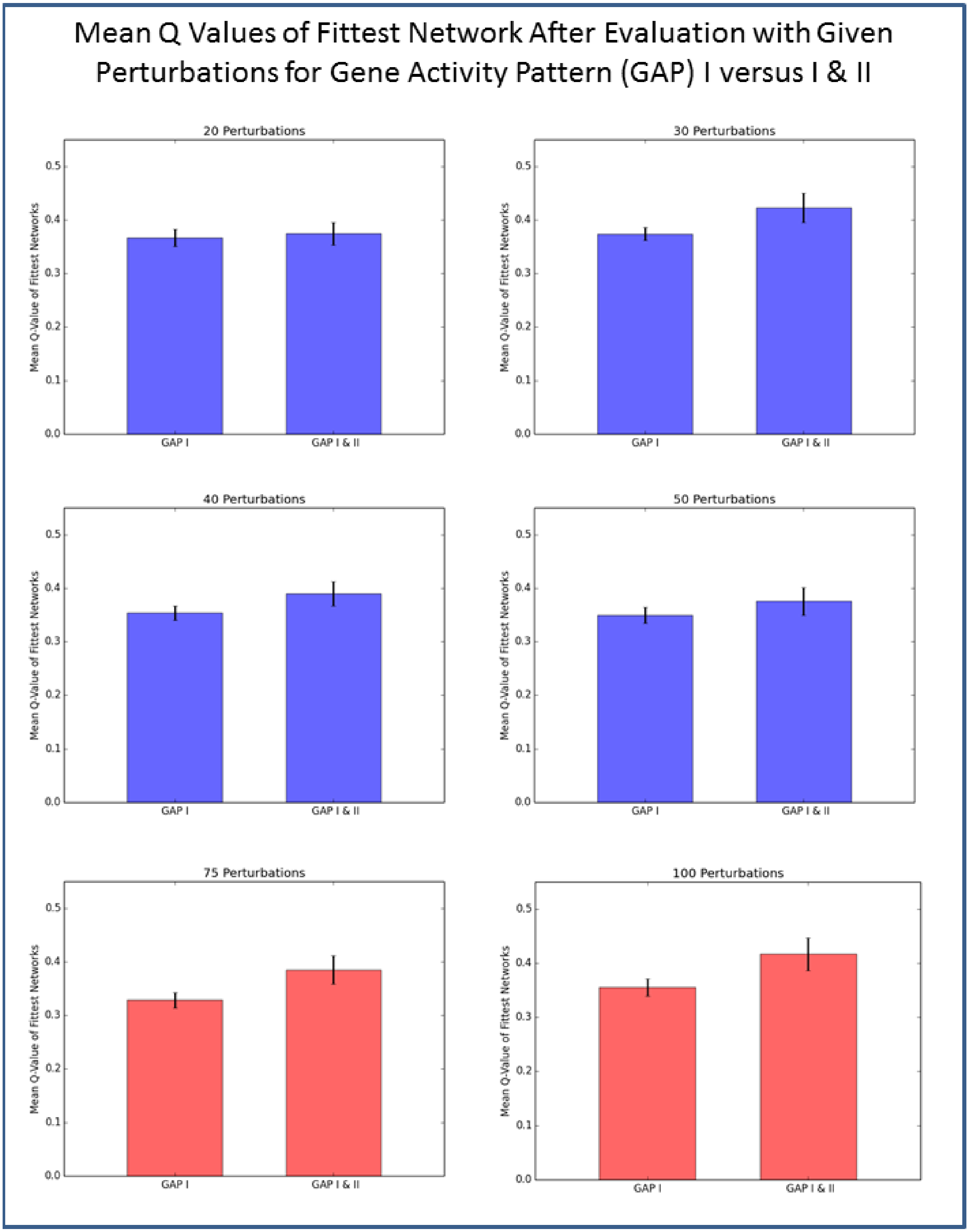


Figure 10. Each of the above graphs shows the mean Q value of the best performing networks over 20 generations after evolution for GAP I and GAP I & II at each level of perturbation. Q was found to be statistically higher after evolving with specialization at 75 and 100 perturbations.

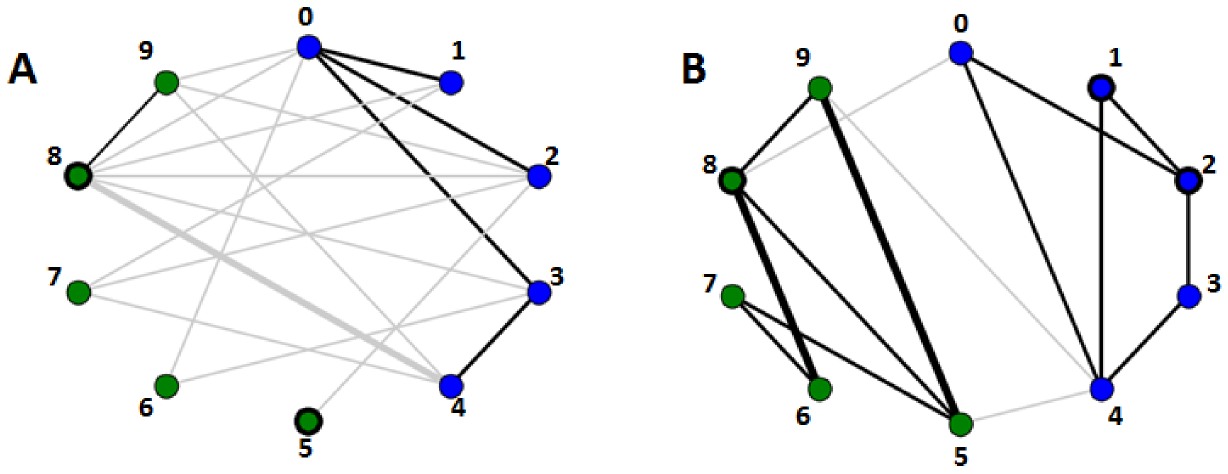


Figure 8. This figure shows two networks that were produced by 2000 generations of evolution and achieved the highest fitness in their population. The modules used in the calculation of Q are indicated by blue and green nodes. Grey edges are used for inter-module interactions and black edges are drawn for intra-module interactions. Bolded edges indicate that gene i influences gene j and gene j also influences gene i. Bolded nodes indicate that gene i has an interaction with itself (self-loop). (A) shows a network that was evaluated with 20 perturbations (Q = 0.192) and (B) shows a network evaluated with 100 perturbations (Q = 0.735). While both networks were able to relatively successfully produce specialized GAPs, network (B) was far more modular than network (A).

4.1b Influence of Biased Mutation on the Evolution of Modularity in

Specialized GRNs. This research also began to investigate how the use of a biased mutation operator drove the evolution of modularity. This biased mutation forced the networks to average 2 – 3 regulators per gene, thus pressuring the networks toward criticality. Initial analysis of GRNs without the biased mutation showed no significant evolution of modularity (see Figure

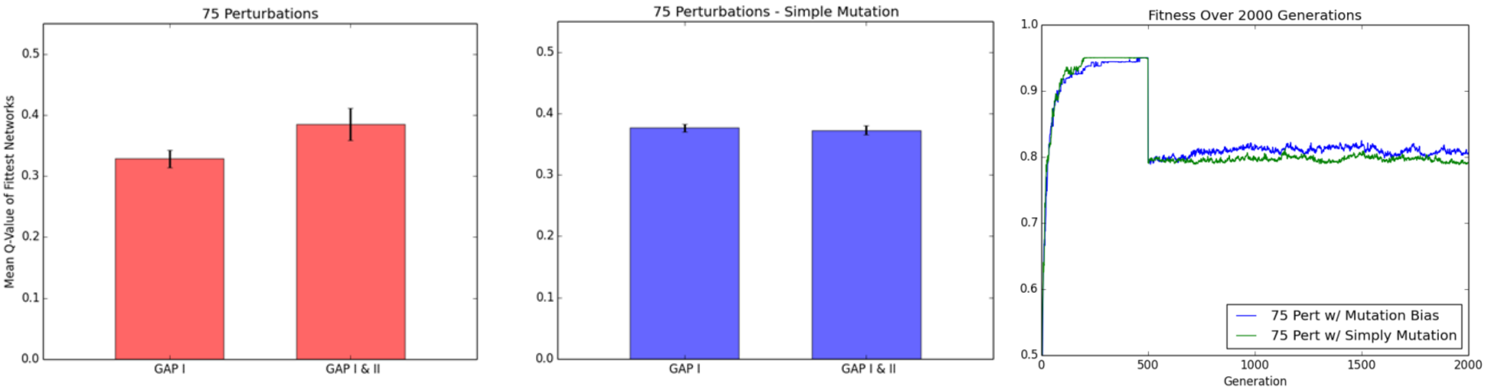


Figure 12. (A) and (B) show the mean Q value of the fittest networks after evolution with and without specialization with 75 perturbations. (A) shows that there was a statistically significant increase in modularity when evolution was performed with a biased mutator. (B) shows the same evolution without a biased mutation and shows no indication of an increase in modularity. (C) shows the fitness of mutation with and without mutation bias with 75 perturbations.

12). This result is likely due to the association between connection cost modularity (Clune et al., 2013). Furthermore, removing the pressure for two to three regulators per gene allows networks to evolve more connections, which causes these networks to behave less consistently.

It should also be noted that simple mutation achieved relatively high Q values giving the impression of high modularity, though not increasing as a result of specialization. This is likely a misrepresentation due to a non-normalized measurement of Q calculated on networks with overall higher numbers of connections.

4.2 Analysis Summary

This study found that 1) modularity evolved in specialized GRNs only when networks are evaluated with 75 or more perturbations 2) without bias mutation modularity does not evolve in specialized GRNs. While there were some limitations to this study, these findings provide opportunity for future research.

4.3 Limitations

While this research provided new insight into the evolution of modularity in GRNs there were several aspects that limited the significance of the results. Namely, time constraints limited the amount and quality of the data that was collected. Given the computational power of the available machine, performing a single run at each level of perturbations required about 20 hours, and a single run with unbiased mutation at 75 perturbations required about 7 hours. For this reason, a total of only 20 runs were used for the statistical analysis. While this sample size is large enough to perform the Student’s t test with 38 degrees of freedom, more runs would have provided a better idea of the distribution of Q values, a more definite idea of the true mean Q values, and stronger idea of statistical significance.

Furthermore, there are likely better measurements of modularity than the Q value used in this research. For one, Q was computed based on an assumed partitioning of the communities into two modules: genes 0-4 and genes 5-9. While these were the hypothesized regions of interest, forcing this assumption onto our measurement may have clouded the true modularity of a network. It would have been more appropriate to use a community detection algorithm, such as the Louvain method, to maximize the measurement of modularity (De Meo, Ferrara, Fiumara, & Provetti, 2011). Additionally, rather than using a raw Q value it would have been more indicative if this Q value had been normalized using the equation $Q_{normal} = (Q – Q_{ran}) / (Q_{max} – Q_{ran})$. Here Q is the maximal Q value of the observed network, Q_{ran} is average Q value of random networks with the same degree distribution of the observed network, and Q_{max} is the highest Q value attained by the group of random networks. Thus Q_{normal} tells how modular a network is relative to random networks with the same attributes. This method of using community maximization and normalizing Q is especially important in measuring modularity of the networks evolved without the biased mutator since these networks tended to be more densely connected.

4.4 Future Research

In addition to addressing the aforementioned limitations it would be interesting to research the effect of sexual recombination and crossover hotspots on the evolution of modularity. Along with random mutation, sexual recombination plays a crucial role in creating variation in the GRN between generations. Recombination refers to “the process of one double-stranded DNA molecule joining with another; specifically in the context of meiosis, the process of two homologous chromosomes exchanging large portions of their DNA (this is called ‘crossing over’),” (Hey, 2004, p. 0730). When a new organism is sexually reproduced, its genome is a composite of splicing together sections of the parents’ DNA.

Within the mass of interconnected genes in the GRN lay “recombination hotspots” (Hey, 2004). These hotspots along the genome indicate locations where rates of breaking and recombining DNA are much higher. The locations of hotspots provide insight into the organization of the new GRN during sexual recombination. Currently, the relationship between cross-over hotspots and modularity has not been studied. However, due to their importance in sexual recombination, it is desirable to include them when modeling the evolution of the GRN. Given that modularity is favored by evolution in specialized GRNs, it is expected that evolution will utilize cross-over hotspots to facilitate and maintain modularity in the GRN.

Like Espinosa-Soto and Wagner (2010), the model used in this research relied entirely on random mutation to create evolutionary variation between generations. Alternatively, during recombination a “child” network could inherit its values from recombining the values of two successful “parent” networks in order to produce the next generation. Additionally, the incorporation of crossover hotspots could be approached in several ways. First, sexual recombination could be implemented by using fitness proportional selection to choose two parent networks and then recombining these networks in a random way to produce a child network. This random form of recombination will provide control data to see if the addition of crossover hotspots facilitates the evolution of modularity. Next, crossover hotspots could be modeled by assigning each population an array, $c^i = (c^0, \ldots , c^{N-2})$, in which an element, c^i, indicates the probability of a crossover occurring between gene i and gene i+1. Elements with a high probability represent crossover hotspots, and the highest probability in the array will dictate where recombination will occur. This array would be evolved along with the networks in a population to determine the most effective position of recombination. Examination of how evolution discovers modularity and where evolution decides to recombine could give significant insight into the relationship between modularity and crossover hotspots.

5. Conclusion

Modeling using mathematics and computer science can provide a unique contribution to biological science. Biology examines a problem with empirical evidence that can be collected and observed from the world. Subsequently, biology is restricted by the tools and technology available to observe and understand the inner workings of organic complex systems. This is especially restrictive in observing the behavior of large networks such as GRNs. Consequently, modeling with computer science and mathematics can be a cost effective and realistic way of studying the GRN. More explicitly, modeling using evolutionary computation has the ability to imitate the natural process of evolution in a semi-controlled and measurable environment. This ties perfectly into modularity because, while modularity has been extensively studied in the biological world, its evolutionary origin is still highly debated. By modeling the evolution of modular biological systems we can gain insight into how modularity might evolve. The results created by these models can then provide new hypotheses that may be tested later with biological experimentation.

As discussed before, modularity in gene regulatory networks is crucial to the development and evolution of life but there is no definite answer for how this modularity arises during evolution. This model of GRNs with evolutionary computation provided new insight into the importance of robustness and criticality in evolving modularity in the GRN. Specifically, we have discovered that networks need to express a significant amount of robustness in order to evolve modularity. Additionally, there is some evidence to suggest that GRN's expression of criticality and maintenance of a relatively sparse network also helps drive the evolution of modularity. This research has provided evidence that these characteristics hold important roles in the behavior and evolution of the GRN, and suggests that both robustness and criticality should continue to be explored in both the computational and observational approaches to biology.

References

Aderem, A. (2005). Systems Biology: Its Practice and Challenges. *Cell*, *121*(4), 511–513. http://doi.org/10.1016/j.cell.2005.04.020

Albert, R., & Othmer, H. G. (2003). The Topology of the Regulatory Interactions Predicts the Expression Pattern of the Segment Polarity Genes in Drosophila Melanogaster. *Journal of Theoretical Biology*, *223*(1), 1–18. http://doi.org/10.1016/S0022-5193(03)00035-3

Callebaut, W., & Rasskin-Gutman, D. (Eds.). (2005). *Modularity: Understanding the Development and Evolution of Natural Complex Systems*. Cambridge, Mass: MIT Press.

Clune, J., Mouret, J.-B., & Lipson, H. (2013). The Evolutionary Origins of Modularity. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1755), 20122863–20122863. http://doi.org/10.1098/rspb.2012.2863

Das, S., Caragea, D., Welch, S., & Hsu, W. H. (Eds.). (2010). *Handbook of Research on Computational Methodologies in Gene Regulatory Networks:*. IGI Global. Retrieved from http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60566-685-3

De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2011). Generalized Louvain method for community detection in large networks (pp. 88–93). IEEE. http://doi.org/10.1109/ISDA.2011.6121636

Eiben, A. E. (2003). *Introduction to Evolutionary Computing*. New York: Springer.

Espinosa-Soto, C., & Wagner, A. (2010). Specialization Can Drive the Evolution of Modularity. *PLoS Computational Biology*, *6*(3), e1000719. http://doi.org/10.1371/journal.pcbi.1000719

Fernández, P., & Solé, R. V. (2003). The Role of Computation in Complex Regulatory Networks. In *Power Laws, Scale-Free Networks and Genome Biology* (pp. 206–225). Boston, MA: Springer US. Retrieved from http://link.springer.com/10.1007/0-387-33916-7_12

Hey, J. (2004). What's So Hot about Recombination Hotspots? *PLoS Biology*, *2*(6), e190. http://doi.org/10.1371/journal.pbio.0020190

Kashtan, N., & Alon, U. (2005). Spontaneous Evolution of Modularity and Network Motifs. *Proceedings of the National Academy of Sciences*, *102*(39), 13773–13778. http://doi.org/10.1073/pnas.0503610102

Lorenz, D. M., Jeng, A., & Deem, M. W. (2011). The Emergence of Modularity in Biological Systems. *Physics of Life Reviews*. http://doi.org/10.1016/j.plrev.2011.02.003

Schlosser, G., & Wagner, G. P. (Eds.). (2004). *Modularity in Development and Evolution*. Chicago: University of Chicago Press.

Shmulevich, I., Dougherty, E. R., Kim, S., & Zhang, W. (2002). Probabilistic Boolean Networks: a Rule-based Uncertainty Model for Gene Regulatory Networks. *Bioinformatics*, *18*(2), 261–274. http://doi.org/10.1093/bioinformatics/18.2.261

Torres-Sosa, C., Huang, S., & Aldana, M. (2012). Criticality Is an Emergent Property of Genetic Networks that Exhibit Evolvability. *PLoS Computational Biology*, *8*(9), e1002669. http://doi.org/10.1371/journal.pcbi.1002669

Wagner, G. P. (1996). Homologues, Natural Kinds and the Evolution of Modularity. *Integrative and Comparative Biology*, *36*(1), 36–43. http://doi.org/10.1093/icb/36.1.36